CRIM250 Final Project

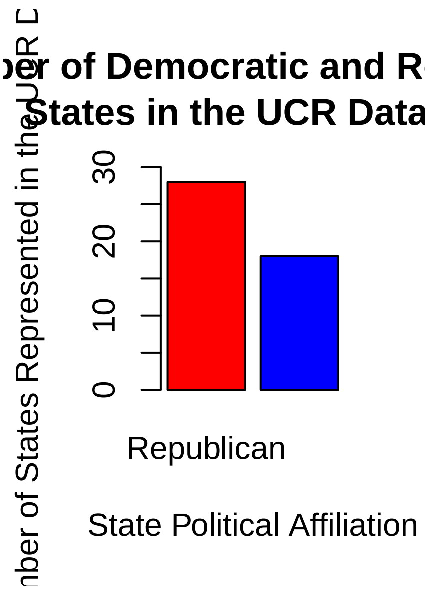
Halle Wasser, Theo Athanitis, and Tori Borlase

Load the data.

```
library(readr)
library(knitr)
dat <- read.csv(file = 'FinalProjectData.csv')
dat5 <- read.csv(file = 'FinalProjectData5.csv')
```

EDA

```
y = data.frame(Political_Leaning=c('Republican', 'Democratic'),Number=c(28,18))
colours = c("red", "blue")
w <- c(0.05, 0.05)
barplot(y$Number, width = w, main='Number of Democratic and Republican
States in the UCR Dataset', ylab='Number of States Represented in the UCR Dataset', xlab='State Politica
```



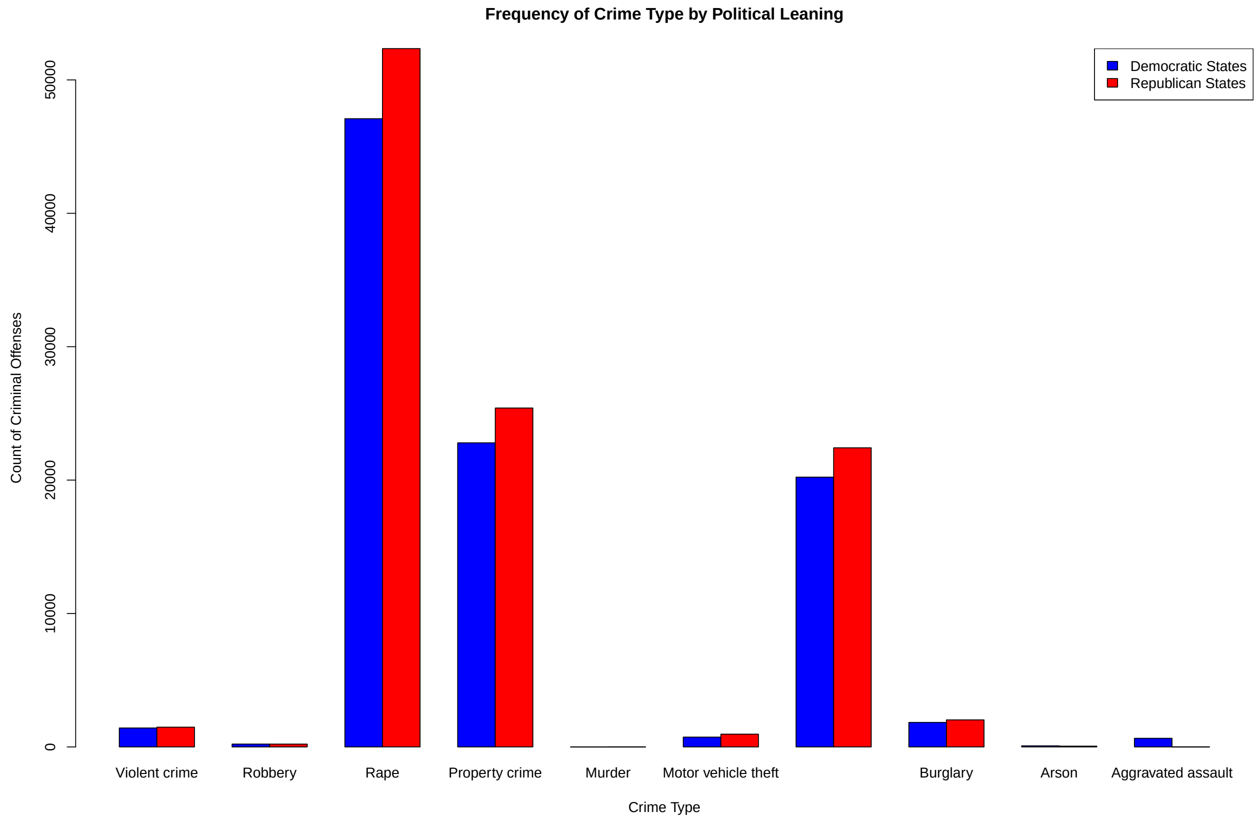
EDA

```
counts5 <- t(as.matrix(dat5[-1]))
counts5
```

```
##    [,1] [,2]  [,3]  [,4] [,5] [,6]  [,7] [,8] [,9] [,10]
## X0 1420  214 47092 22796    2  736 20223 1837   82   646
## X1 1481  211 52346 25404    5  957 22423 2028   58     5
```

```
colnames(counts5) <- dat5$crime_type
colours = c("blue", "red")
barplot(counts5, main='Frequency of Crime Type by Political Leaning', ylab='Count of Criminal Offenses'
        col=colours, ylim=c(0,max(counts5)*1))
```

```r
legend('topright',fill=colours,legend=c('Democratic States','Republican States'))
```



Linear Regressions

```r
# Correlation between Crime and Political Affiliation
cor(dat$State.Leaning, dat$Total)
```

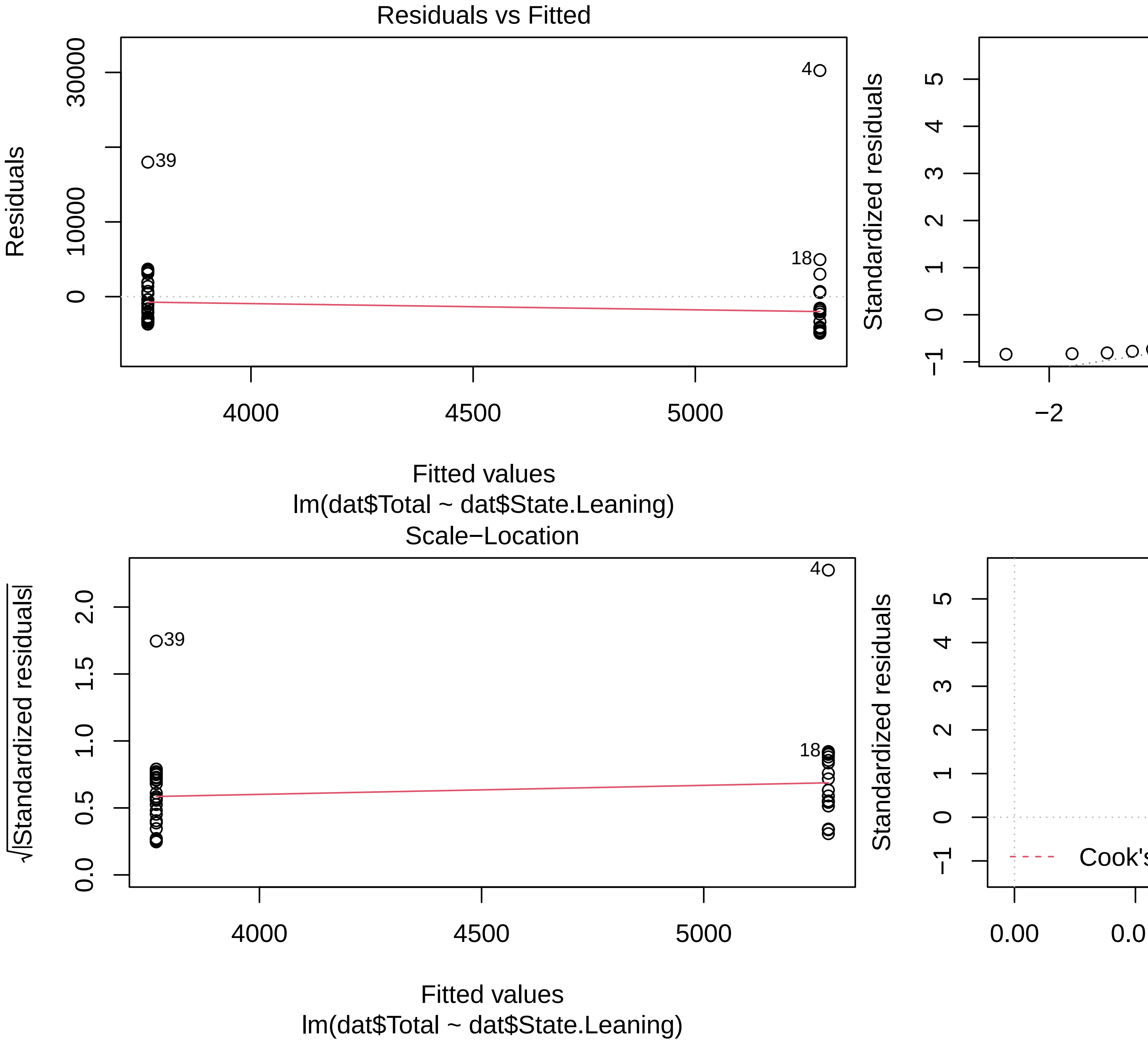
```
## [1] -0.1245639
```

```r
# Total Crime Regression
reg.output <- lm(dat$Total ~ dat$State.Leaning, data = dat)
summary(reg.output)
```

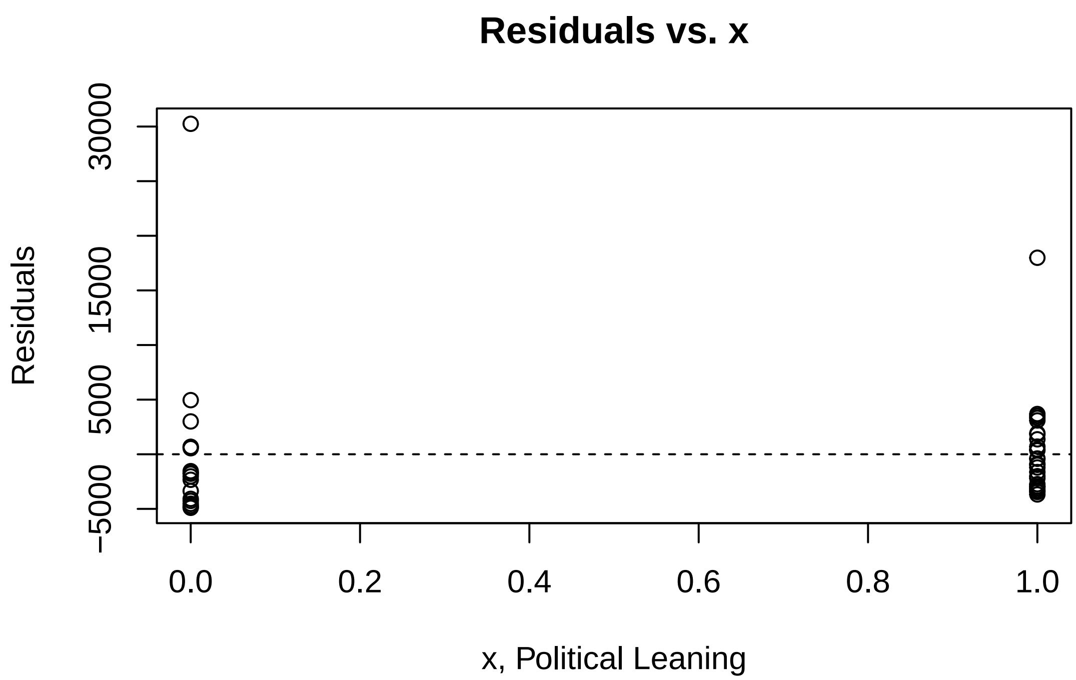
```
##
## Call:
## lm(formula = dat$Total ~ dat$State.Leaning, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4901.4 -3097.1 -1587.2   698.5 30250.6
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5280       1417   3.726 0.000551 ***
## dat$State.Leaning    -1513       1816  -0.833 0.409484
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6012 on 44 degrees of freedom
```

```
## Multiple R-squared:  0.01552,	Adjusted R-squared:  -0.006858
## F-statistic: 0.6935 on 1 and 44 DF,  p-value: 0.4095
```

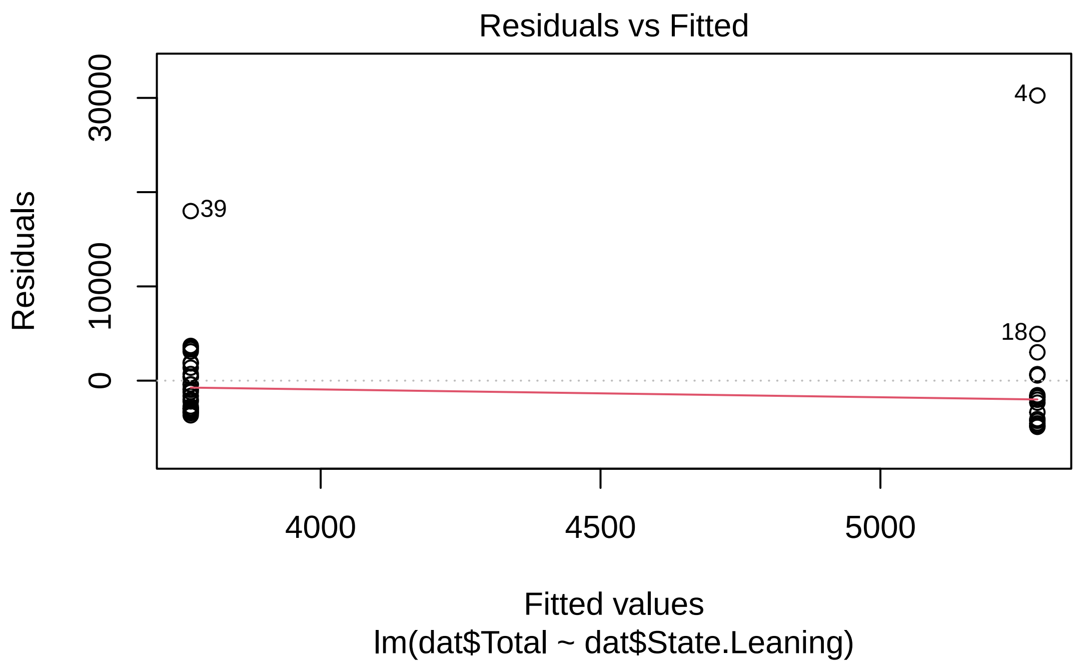
```
plot(reg.output)
```



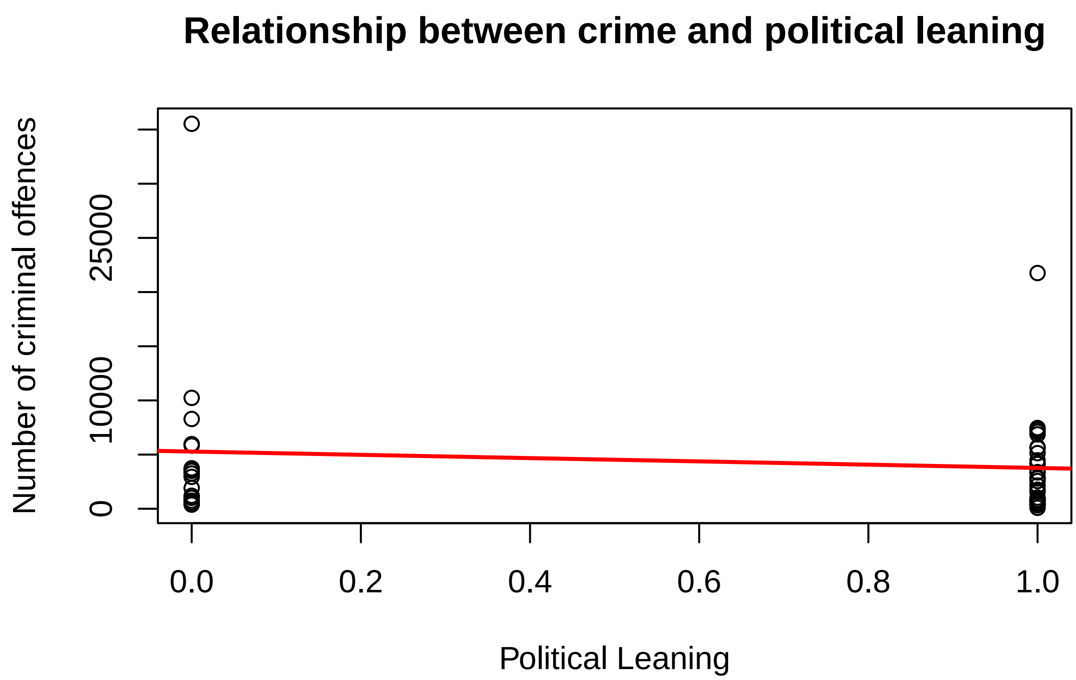

```
# Linearity Assumption:
plot(dat$State.Leaning, reg.output$residuals, main="Residuals vs. x", xlab="x, Political Leaning", ylab=
abline(h = 0, lty="dashed")
```



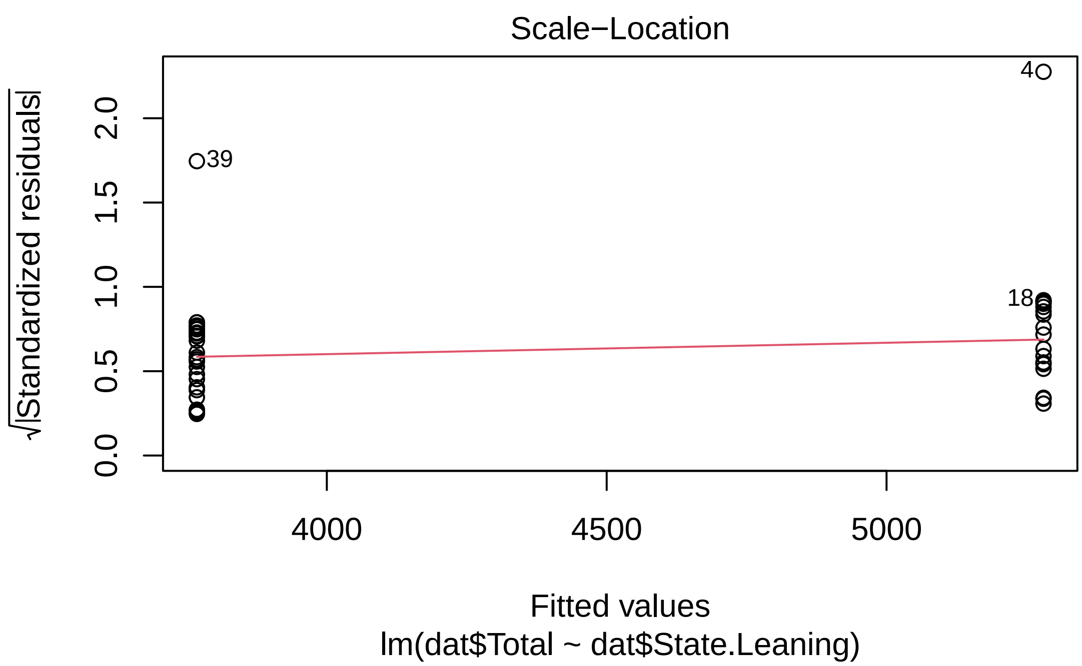

```
plot(reg.output, which=1)
```



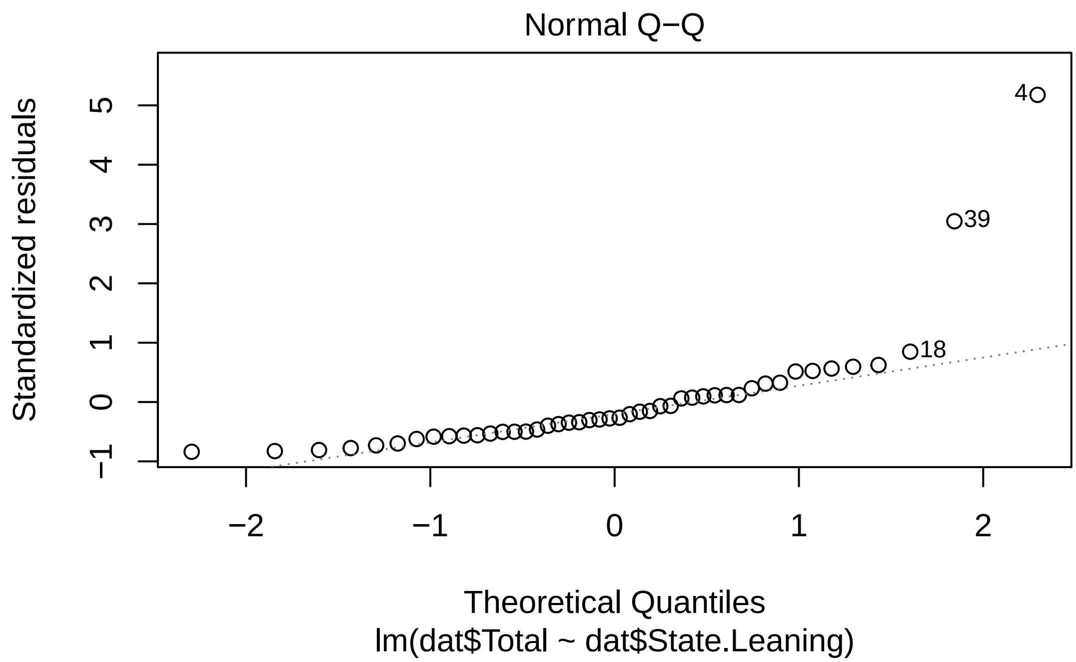

```
# Independence Assumption: using Residuals vs. x plotted above
plot(dat$State.Leaning, dat$Total, main="Relationship between crime and political leaning",
    xlab="Political Leaning", ylab="Number of criminal offences")
abline(reg.output, col = "red", lwd=2)
```



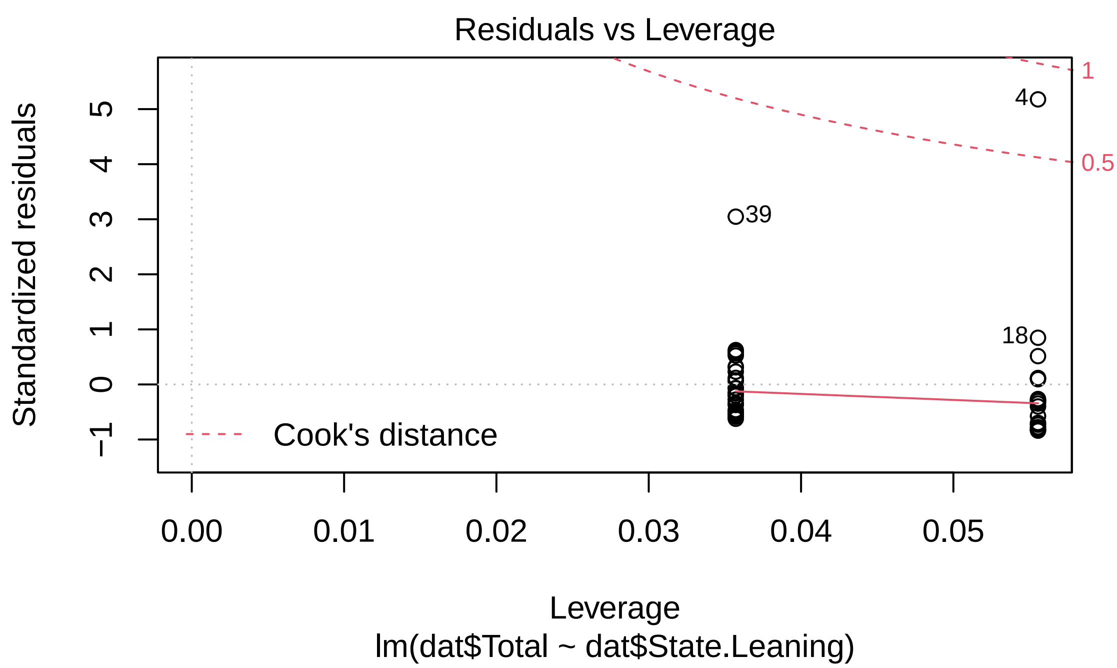

```
# Equal Variance Assumption/ Homoscedasticity:
plot(reg.output, which=3)
```




```
# Normal Population Assumption:
plot(reg.output, which=2)
```




```
plot(reg.output, which=5)
```




```
# Rape regression
reg.output1 <- lm(dat$Rape ~ dat$State.Leaning, data = dat)
summary(reg.output1)
```

```
##
## Call:
## lm(formula = dat$Rape ~ dat$State.Leaning, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -2427.2 -1536.8  -780.4   348.2 14986.8
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2616.2      702.8   3.723 0.000557 ***
## dat$State.Leaning    -746.7      900.8  -0.829 0.411605
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2982 on 44 degrees of freedom
## Multiple R-squared:  0.01538,    Adjusted R-squared:  -0.007001
## F-statistic: 0.6872 on 1 and 44 DF,  p-value: 0.4116
```

```
# Violent Crime Regression
reg.output2 <- lm(dat$Violent.crime ~ dat$State.Leaning, data = dat)
summary(reg.output2)
```

```
##
## Call:
## lm(formula = dat$Violent.crime ~ dat$State.Leaning, data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -77.89 -46.64 -19.39  31.11 417.11
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)           78.89      19.20   4.108 0.000171 ***
## dat$State.Leaning    -26.00      24.62  -1.056 0.296690
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 81.48 on 44 degrees of freedom
## Multiple R-squared:  0.02472,    Adjusted R-squared:  0.002556
## F-statistic: 1.115 on 1 and 44 DF,  p-value: 0.2967
```

Linearity Assumption: This assumption is met. The residuals vs. x plot has a horizontal direction and does have a significant pattern in the data. Furthermore, the residuals vs fitted plot is fairly horizontal and flat, meaning that there is no discernible non-linear trend to the residuals.

Independence Assumption: This assumption is also met for the same reason as the linearity assumption as the residuals vs. x plot has a horizontal direction and does have a significant pattern in the data, as well as because there does not seem to be a time-series component to the data.

Equal Variance Assumption/ Homoscedasticity: This assumption is not met. The scatter plot of crimes vs. political affiliation has no variations with shrinkage in the plot. Additionally, there are significant negative trends, based on the size of the data, shown by the line in the scale-location plot showing that the errors do not have a constant variance.

Normal Population Assumption: This assumption is not met as the q-q plot has significant left skew deviations and heavy-tailed for values in this plot.

As these assumptions are not met, normally the next step would be to use the box-cox method to find the best transformation for this data, transform the x variable, and repeat this process. However, as the p-value was so large at 0.4095, demonstrating that this relationship is not statistically significant, we instead concluded that we cannot reject the null hypothesis and instead explored whether or not this relationship existed for a particular crime type variable. However, the two variables that we explored, rape and violent crime, also had significant p-values of 0.4116 and 0.2967 respectively and we concluded that we cannot reject the null hypothesis for these indivigual variables either. While this analysis does not show a relationship between crime frequency on college campuses and the political affiliation of the state in which it is located, in the following section we will argue that it may actually exist based on a series of confounding factors and that the limitations of this dataset make it impossible to distinguish in this analysis.